

# Taenia Biomolecular Phylogeny and the Impact of Mitochondrial Genes on this Latter

Huda Al-Nayyef<sup>1,2</sup>, Christophe Guyeux<sup>1</sup>, and Jacques M. Bahi<sup>1</sup>

<sup>1</sup> FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department  
Université de Franche-Comté, 16, Rue de Gray, 25000 Besançon, France

<sup>2</sup> Computer Science Department, University of Mustansiriyah, Iraq  
{huda.al-nayyef, christophe.guyeux, jacques.bahi}@univ-fcomte.fr

**Abstract**—Variations in mitochondrial genes are usually considered to infer phylogenies. However some of these genes are lesser constraint than other ones, and thus may blur the phylogenetic signals shared by the majority of the mitochondrial DNA sequences. To investigate such effects, in this research work, the molecular phylogeny of the genus *Taenia* is studied using 14 coding sequences extracted from mitochondrial genomes of 17 species. We constructed 16,384 trees, using a combination of 1 up to 14 genes. We obtained 131 topologies, and we showed that only four particular instances were relevant. Using further statistical investigations, we then extracted a particular topology, which displays more robustness properties.

**Index Terms**—*Taenia*, Phylogeny, Statistical tests

## I. INTRODUCTION

*Taenia* (Cestoda: Taeniidae) is a genus of tapeworm (a type of helminth) members that have some important parasites of livestock. These parasitic organisms handle taeniasis and cysticercosis in humans, which are a type of helminthiasis that was belonging to the group of neglected tropical diseases [22]. Despite intensive research, the taxonomy of this genus remains unclear. Based on morphology and life cycle data. An essential key to solve the last issues raised by *Taenia* is believed to be found in the study of the large amount of recently available DNA sequences, especially with complete mitochondrial (mt) genomes. Genes from mt genomes are classical markers for phylogeny. This DNA presents interesting features for such analysis: genes are shared by almost all eukaryotes and are present in a single copy, the molecule is maternally inherited and non-recombining in most cases, etc. [4].

Part of the problem resides in the fact that, even though the amount of information should be sufficient to infer a correct phylogeny of this genus. The presence of homoplasmy in individually available genes clouds the general phylogenetic message, raising uncertainties in some locations of the tree. The question we discuss in the present work is thus to determine which genes are homoplastic, and which ones tell the story of the species. Our goal is thus to exhibit a well-supported phylogenetic tree of the genus *Taenia*. Our analysis relies on some recent statistical tools and intensive computations on available bio-molecular data [2], [3].

After a presentation of the major problems that remain need to be solved regarding the phylogeny of *Taenia*, we will describe in details our investigation protocol. We will then present how each phylogenetic tree inference has been

TABLE I: *Taenia* species analyzed in this paper and accession numbers of mitochondrial genomes. (*E. vogeli*) is an outgroup.

Species	Accession
<i>Taenia asiatica</i>	NC_004826
<i>Taenia crassiceps</i>	NC_002547
<i>Taenia hydatigena</i>	NC_012896
<i>Taenia krepkogorski</i>	NC_021142
<i>Taenia laticollis</i>	NC_021140
<i>Taenia madoquae</i>	NC_021139
<i>Taenia martis</i>	NC_020153
<i>Taenia multiceps</i>	NC_012894
<i>Taenia mustelae</i>	NC_021143
<i>Taenia ovis</i>	NC_021138
<i>Taenia parva</i>	NC_021141
<i>Taenia pisiformis</i>	NC_013844
<i>Taenia saginata</i>	NC_009938
<i>Taenia serialis</i>	NC_021457
<i>Taenia solium</i>	NC_004022
<i>Taenia taeniaeformis</i>	NC_014768
<i>Taenia twitchelli</i>	NC_021093
<i>Echinococcus vogeli</i>	NC_009462

conducted. Our approach mainly based on annotating from scratch each genome, using an efficient alignment tool, and various mutation models for mitochondrial coding sequences and RNAs. We will then explain how we have obtained the 16,384 phylogenetic trees of that study, and how we have used them to solve the phylogenetic reconstruction problem for this genus as a result of estimating the influence of each gene on that topology.

To date, 17 complete mitochondrial genomes of *Taenia* have been published, their list and accession number being provided in Table I. These genomes have been used recently to update the phylogeny of this genus using molecular data. As presented in Table II many previous articles of phylogeny have worked with *Taenia* species, but none of them provides a well-supported tree for this genus. For this reason, the authors of this paper proposed the new computational methods for constructing and finding a well-supported phylogenetic tree for *Taenia* [1].

The remainder of this article is constituted as follows. Section II is devoted to the proposed methodology intended to improve the estimation of the phylogenetic tree. Finer statistical investigations of the homoplastic character of certain genes are detailed in Section III. This article ends with a discussion and a description of possible future work on this problem.

TABLE II: *Taenia (Eucestoda)* in state of the art phylogenies (\* when serving as outgroup; + when present in dataset but not used in phylogenetic analyses;  $X^n$  when  $n$  represents of the species).

	Hoberg <i>et al.</i> 2000 [12]	Hoberg <i>et al.</i> 2001 [13]	Hoberg 2006 [11]	Nakao <i>et al.</i> 2010 [18]	Lavikainen <i>et al.</i> 2010 <sup>o</sup> [16]	Knapp <i>et al.</i> 2011 <sup>o</sup> [14]	Nakao <i>et al.</i> 2013 <sup>o</sup> [17]	This study <sup>o</sup>	Total of studies
<i>Taenia acinonyxi</i>	X	X	X						3
<i>Taenia asiatica</i>	X	X	X	X	X	X	X	X	8
<i>Taenia brachyacantha</i>	X	X							2
<i>Taenia crassiceps</i>	X	X	X	X	X	X	X	X	8
<i>Taenia crocutae</i>	X	X	X	X					4
<i>Taenia dinniki</i>	X	X	X						2
<i>Taenia endotheracicus</i>	X	X	X						3
<i>Taenia gonyamai</i>	X	X	X						3
<i>Taenia hyaenae</i>	X	X	X	X					4
<i>Taenia hydatigena</i>		X	X		X	X	X	X	6
<i>Taenia ingwei</i>	X	X							2
<i>Taenia intermedia</i>			X						1
<i>Taenia krabbei</i>					X				1
<i>Taenia krepkogorski</i>							X	X	2
<i>Taenia laticollis</i>	X	X				X	X	X	5
<i>Taenia macrocystis</i>	X	X	X						3
<i>Taenia madoquae</i>	X	X	X		X	X	X	X	7
<i>Taenia martis</i>	X	X	X		X	X	X	X	7
<i>Taenia multiceps</i>	X	X	X	X	X	X	X	X	8
<i>Taenia mustelae</i>	X	X	X		X	X	X	X	7
<i>Taenia olngojinei</i>	X	X	X						3
<i>Taenia omissa</i>	X	X	X						3
<i>Taenia ovis</i>	X	X	X <sup>2</sup>		X	X	X	X	7
<i>Taenia parenchymatosa</i>	X	X	X <sup>3</sup>						3
<i>Taenia parva</i>	X	X	X		X	X	X	X	7
<i>Taenia pencei</i>			X						1
<i>Taenia pistiformis</i>	X	X	X		X			X	5
<i>Taenia polyachantha</i>	X	X			X <sup>2</sup>				3
<i>Taenia pseudolaticollis</i>	X	X							2
<i>Taenia regis</i>	X	X	X		X				4
<i>Taenia rileyi</i>	X	X	X						3
<i>Taenia saginata</i>	X	X	X	X	X	X	X	X	8
<i>Taenia selousi</i>	X	X	X						3
<i>Taenia serialis</i>	X	X	X <sup>2</sup>	X <sup>2</sup>	X	X	X	X	8
<i>Taenia simbae</i>	X	X	X	X					4
<i>Taenia solium</i>	X	X	X	X	X	X	X	X	8
<i>Taenia taeniaeformis</i>	X	X	X	X	X <sup>2</sup>	X	X	X	8
<i>Taenia taxidiensis</i>	X	X	X						3
<i>Taenia twitchelli</i>	X	X	X		X	X	X	X	7
<i>Echinococcus vogeli</i>				X		X	+	*	3
Total 40	34	35	31	11	18	16	16	17	

## II. MATERIALS AND METHODS

### A. Alignment and annotations of coding sequences

To answer the aforementioned questions, first Bayesian and maximum likelihood analyses have been realized on either the whole mitogenomes or its twelve protein coding genes. These analyses were realized using nucleotides and translated amino acids sequences. Tools used during these first runs of analyses were:

- Muscle [6] for aligning complete mitogenomes and T-Coffee [19] for genes alignments;
- NCBI annotations for coding sequences in a first analysis, and then DOGMA [24] in a deeper stage;
- PhyloBayes [15] for Bayesian inference, while PhyML [9] and RAxML [23] have been used for maximum likelihood.

TABLE III: Details of obtained topologies. The lowest bootstrap and the number of occurrence of each calculated topology is indicated.

Topology	Lowest bootstrap	Number of occurrences	Average bootstrap	Discarded genes
0	82	2049	44	<i>Atp6, Cob, Cox2, Nad1, Nad2, Nad3, Nad5</i>
1	84	6442	51	<i>Nad1, Nad3, Nad5, Nad6, Rrms</i>
2	92	3276	52	<i>Cox2, Cox3, Nad4, Nad4l, Nad5, Rrnl, Rrms</i>
3	76	931	48	<i>Atp6, Cox1, Nad1, Nad3, Nad4, Rrnl</i>
4	74	452	52	<i>Atp6, Cob, Cox1, Cox3, Nad4, Nad5, Rrnl</i>
5	56	317	28	<i>Cob, Cox1, Cox2, Cox3, Nad1, Nad2, Nad3, Nad4l, Rrnl, Rrms</i>
6	68	614	39	<i>Atp6, Cox1, Cox2, Cox3, Nad2, Nad3, Nad5</i>
7	68	321	43	<i>Atp6, Cox2, Cox3, Nad1, Nad2, Nad3, Nad4, Nad4l, Nad6, Rrms</i>
8	70	226	46	<i>Cob, Cox1, Cox2, Cox3, Nad4, Nad4l</i>
9	58	69	39	<i>Cox1, Cox2, Cox3, Nad1, Nad3, Nad4, Rrms</i>
10	74	230	45	<i>Atp6, Cob, Cox1, Nad1, Nad2, Nad4, Nad4l, Nad6, Rrnl</i>
11	76	172	53	<i>Cob, Cox1, Cox2, Cox3, Nad1, Nad3, Nad4, Nad5, Rrnl</i>
12	60	212	30	<i>Atp6, Cox2, Cox3, Nad1, Nad2, Nad4l, Nad6, Rrms</i>
13	56	92	42	<i>Atp6, Cob, Cox1, Cox2, Cox3, Nad1, Nad3, Nad4</i>
14	64	39	44	<i>Atp6, Cob, Cox1, Cox2, Nad3, Nad4, Nad5, Nad6, Rrms</i>

At each time, a problem of support (at least one bootstrap lower than 95, while a commonly accepted rule claims that all supports must be larger than this threshold [7]) was found at least at one location of the obtained tree. Partial conclusions of these preliminary studies were that: (1) to use coding sequences is better than to consider the whole mitogenome, (2) there are inconsistencies in NCBI annotations, (3) T-Coffee alignments seem better than muscle ones, (4) many coding sequences narrate the story of the genus while others tell their own history, and (5) to enlarge the amount of data leads to more supported trees.

### B. Methodological approach

To solve both the phylogeny of *Taenia* and the determination of genes that break it, a solution has been to consider all available or obtainable coding sequences shared by these 18 species, and to investigate how the inferred phylogenies evolve when using a various subset of these sequences. Doing so enlarge the first investigations of Hardman *et al.* [10], who have studied the phylogeny of 5 *Taeniidae* according to each of the 12 mitochondrial genes taken alone, 14 sequences have been extracted from each of the considered species: 12 protein coding sequences and 2 rRNAs from the mitochondrial genomes. They are listed below.

- Mitochondrial protein coding sequences:
  - atp6* (ATP synthase 6), *cob* (cytochrome b), *cox1* (cytochrome c oxydase 1), *cox2* (cytochrome c oxydase 2), *cox3* (cytochrome c oxydase 3), *nad1* (NADH dehydrogenase subunit 1), *nad2* (NADH dehydrogenase subunit 2), *nad3* (NADH dehydrogenase subunit 3), *nad4* (NADH dehydrogenase 4), *nad4l* (NADH dehydrogenase subunit 4L), *nad5* (NADH dehydrogenase subunit 5), *nad6* (NADH dehydrogenase subunit 6).
- Mitochondrial rRNAs:
  - rrnL* (large subunit rRNA), *rrnS* (small subunit rRNA).

DOGMA, for its part, has been used to annotate from scratch each up-to-date complete mitochondrial genome downloaded from NCBI [5] Default parameters of DOGMA have been selected, namely an identity cutoff for protein equal to 60% and 80% for coding genes and rRNAs respectively for *Taenia* species, while these thresholds have been reduced to 55% and 75% for *T. mustelae*, due to a problem of detection of *nad6*

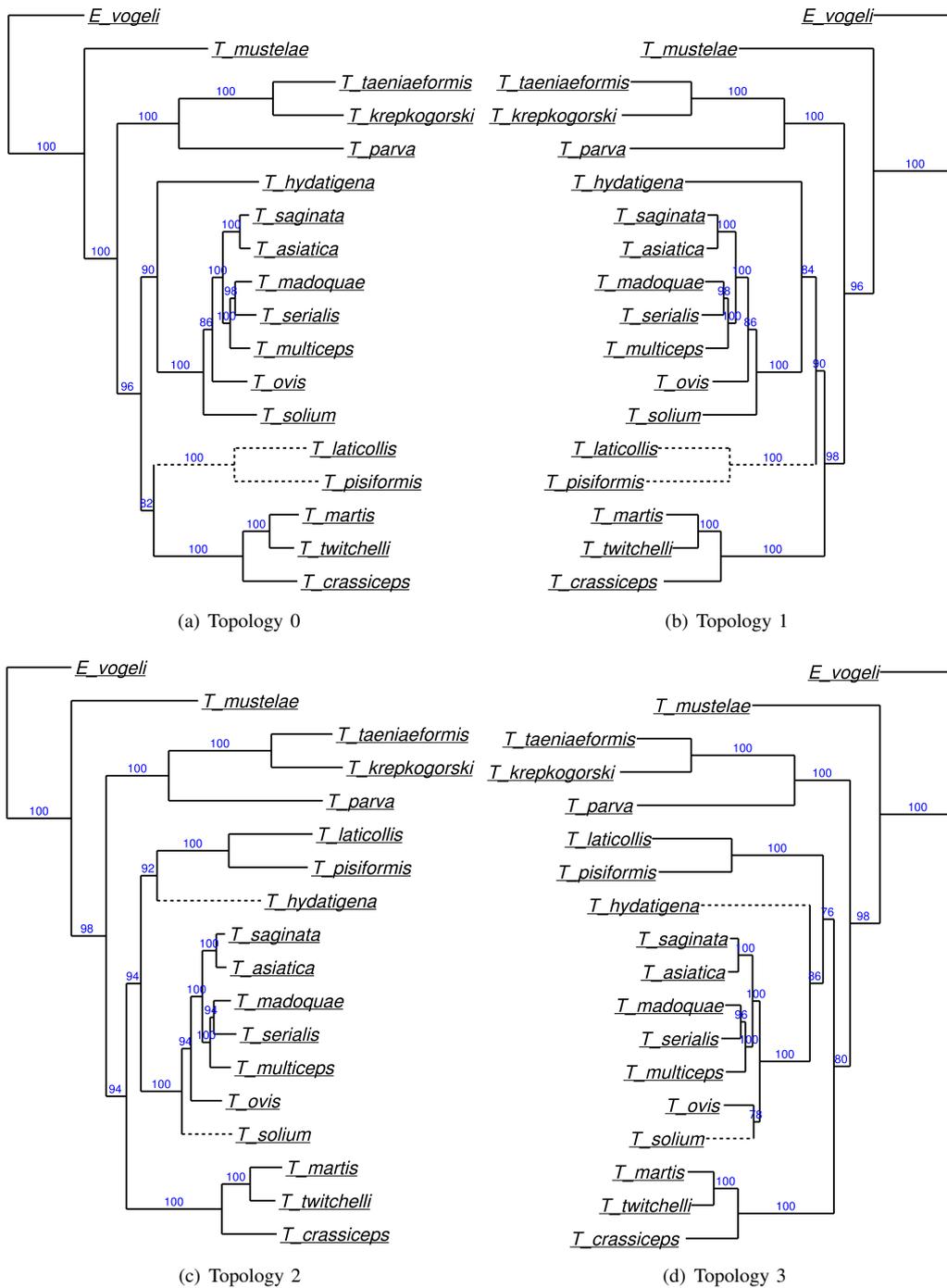


Fig. 1: 4 trees with more than 700 occurrences, when considering 16,384 trees obtained with Algorithm 1 (*E. vogeli* is an outgroup)

and *rrnL* respectively. The e-value was equal to  $1e - 5$ , and the number of blast hits to return has been set to 5.

Each of these 14 coding sequences has been aligned separately by using T-Coffee (M-Coffee mode, using 6 cores for multiprocessing). Then 16,384 trees were constructed, corresponding to all the possible combinations of 1, 2, 3, ..., and 14 coding sequences among the 14 available ones ( $\sum_{k=1}^{14} \binom{14}{k} = 16,384$ ), as described in Algorithm 1. This computation has taken 3 months on the ‘‘Mésocentre de Calcul de Franche-Comté’’ supercomputer facilities. The idea behind was to determine both the most supported phylogenetic trees and the effects of each gene on topologies and supports. RAxML version 8.0.20 were used for maximum likelihood inference, with 3 distinct models/data partitions with joint branch length optimization at each computation, corresponding to the mitochondrial rRNAs, and the mitochondrial protein coding sequences. All free model parameters have been estimated by RAxML for both GAMMA model of rate heterogeneity and ML estimate of alpha-parameter. At each time, a maximum of 1000 non-parametric bootstrap inferences was executed, with MRE-based bootstrapping criterion, and *E. vogeli* has been used as outgroup.

```

for  $k=1, \dots, 14$  do
  for each combination c of k genes do
    build a phylogenetic tree T based on these k
    genes;
    extract the list of bootstraps L(c) and the
    topology T(c);
    store (c, L(c), T(c));
  end
end

```

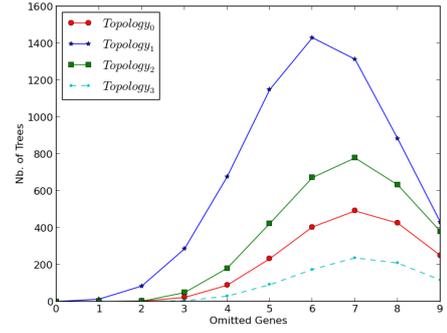
**Algorithm 1:** Pseudocode producing 16,384 phylogenetic trees

### III. DISCUSSION AND RESULTS

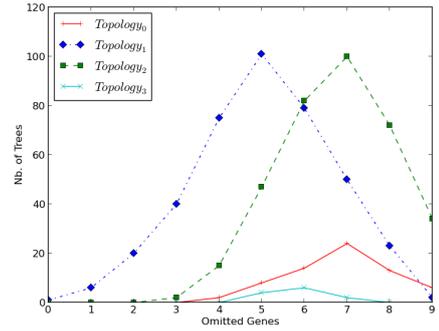
#### A. Results

131 topologies were obtained during our computations with 17 species and 1 outgroup. Further information regarding these trees are provided in Table III: in this latter, we investigated the 15 most frequent topologies that contained 15,442 of the 16,384 trees (94.16%). For each topology, the lowest bootstrap of the best tree (that is, the lowest bootstrap of the tree that maximizes the minimum taken over all its bootstraps), the number of trees having this topology, the average minimal bootstrap value, and the list of genes that have been removed to obtain the best tree having this topology, are provided. Only 4 of these 131 topologies have a number of occurrences larger than 700, when considering the 16,384 obtained trees. They are depicted in Figure 1.

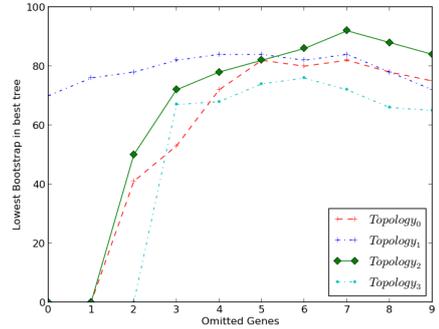
These 4 best topologies representing 77.07% of the obtained trees share most of their structure. For instance, *T. madoquae*, *T. serialis*, *T. multiceps*, are within a same clade, which is sister to the clade consisting of *T. asiatica* and *T. saginata*. The differences between these most frequent topologies are depicted with dotted lines in Figure 1 while Table IV summarizes them using CompPhy tool [8].



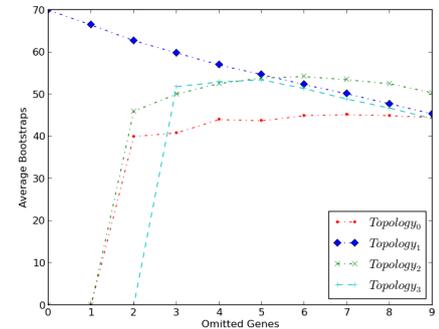
(a) Number of trees per topology.



(b) Number of trees whose lowest bootstrap is larger than 70.



(c) Lowest bootstrap in best trees.



(d) Average value of lowest bootstraps

Fig. 2: Comparison of the 4 best topologies, according to the number of discarded genes. A. The number of trees in each topology, according to the number of discarded genes. B. As in (a), but by considering only trees whose supports are larger than 70. C. Minimal support in the best tree of each topology, when regarding the number of discarded genes. D. Average of all minimum bootstrap in each tree of each topology.



```

for each gene  $g$  do
  for each topology  $t$  do
    count[ $g$ ][ $t$ ] = 0;
  end
end
for each  $(c, L, T)$  in the list stored by Algorithm 1 do
  for each  $g$  in  $c$  do
    count[ $g$ ][ $t$ ] = count[ $g$ ][ $t$ ]+1;
  end
end

```

**Algorithm 2:** Pseudocode producing Table V

TABLE V: Number of times each gene were present to produce a tree having one of the 4 most relevant topologies.

Topologies	0		1		2		3	
	number	rank	number	rank	number	rank	number	rank
<i>atp6</i>	924	9	3431	8	1924	4	423	9
<i>cob</i>	787	11	<b>4179</b>	<b>2</b>	1691	6	350	11
<i>cox1</i>	1209	4	<b>4324</b>	<b>1</b>	1326	11	542	5
<i>cox2</i>	1152	6	3740	6	1472	8	<b>686</b>	<b>2</b>
<i>cox3</i>	1469	3	3966	4	<b>1260</b>	<b>13</b>	449	7
<i>nad1</i>	<b>584</b>	<b>13</b>	3379	12	<b>2549</b>	<b>2</b>	257	12
<i>nad2</i>	<b>84</b>	<b>14</b>	3391	11	<b>3004</b>	<b>1</b>	527	6
<i>nad3</i>	1142	7	<b>2708</b>	<b>13</b>	2069	3	448	8
<i>nad4</i>	<b>1699</b>	<b>1</b>	3677	7	1339	10	<b>84</b>	<b>13</b>
<i>nad4l</i>	1153	5	3421	10	1291	12	624	3
<i>nad5</i>	613	12	4139	3	<b>694</b>	<b>14</b>	<b>883</b>	<b>1</b>
<i>nad6</i>	937	8	3421	9	1887	5	390	10
<i>rrnL</i>	858	10	3855	5	1583	7	<b>63</b>	<b>14</b>
<i>rrnS</i>	<b>1638</b>	<b>2</b>	<b>2598</b>	<b>14</b>	1392	9	603	4

for Topology 2 (*i.e.*, this mitochondrial coding sequence gene plays an essential role in Topology 2).

- 2) It seems that taking *nad5* into consideration leads to a move of *T. hydatigena* in the tree, as it is ranked 12 and 14 for Topologies 0 and 2 respectively, and 3 and 1 for Topologies 1 and 3 respectively.
- 3) Similarly, *rrnS* seems responsible for the position evolution of *T. laticollis* and *T. pisiformis*: this gene is ranked in 2nd position for Topology 0 while this is the least frequent gene (last position) for Topology 1.
- 4) Gene *nad5* is ranked first for Topology 3, so it may impact the sister relationship between *T. solium* and *T. ovis*.

However, these claims need to be further investigated by a more rigorous statistical approach, which is the aim of the following sections.

### C. Genes influence on topology using Dummy logit model

To investigate more deeply the effects of each coding sequence on the species topology, 4 dummy binary choice logit models have been realized (one per each best topology) using `scikit-learn` [20] module of Python language. The reference to the exogenous design is a  $14 \times 16,384$  array, each row being a vector of 0's and 1's: a 0 in position  $i$  of row  $k$  means that, in the  $k$ -th tree computation, gene number  $i$  (in alphabetic order) were discarded, and conversely it was considered if the coefficient is 1. Rows are thus the "observations" while columns correspond to regressors. The 1-d endogenous response variable, for its part, was a vector of size 16,384, having an 1 in position  $k$  if and only if Topology 1 has been produced with the choice of genes corresponding to

TABLE VI: Dummy logit regression results for Topology 1

	coef	std err	z	$P >  z $	[95.0% Conf. Int.]
<i>atp6</i>	-0.2412	0.034	-7.06	0.000	[-0.308, -0.174]
<i>cob</i>	0.6861	0.035	19.871	0.000	[0.618, 0.754]
<i>cox1</i>	0.8592	0.035	24.733	0.000	[0.791, 0.927]
<i>cox2</i>	0.1444	0.034	4.231	0.000	[0.078, 0.211]
<i>cox3</i>	0.4261	0.034	12.431	0.000	[0.359, 0.493]
<i>nad1</i>	-0.3059	0.034	-8.944	0.000	[-0.373, -0.239]
<i>nad2</i>	-0.2915	0.034	-8.526	0.000	[-0.359, -0.224]
<i>nad3</i>	<b>-1.1113</b>	0.035	-31.673	0.000	[-1.18, -1.042]
<i>nad4</i>	0.0658	0.034	1.928	0.054	[-0.001, 0.133]
<i>nad4l</i>	-0.2532	0.034	-7.409	0.000	[-0.32, -0.186]
<i>nad5</i>	0.6381	0.034	18.512	0.000	[0.571, 0.706]
<i>nad6</i>	-0.2537	0.034	-7.423	0.000	[-0.321, -0.187]
<i>rrnL</i>	0.2873	0.034	8.403	0.000	[0.22, 0.354]
<i>rrnS</i>	<b>-1.2345</b>	0.035	-35.003	0.000	[-1.304, -1.165]

the row number  $k$  in the exogenous design (resp. Topology 0, 2, or 3 in the three other binary choice logit models). The model has been fitted using maximum likelihood with Newton-Raphson solver. Convergence has been obtained after 8 iterations, and the Logit regression results are summarized in Table VI.

A first conclusion of the results obtained when investigating the impact of each gene on the most supported topology is that all considered coding sequences bring information, except perhaps the particular case of *nad4* (see column  $P > |z|$ ). Additionally, when the effect of a mitochondrial coding sequence is negative regarding Topology 1, its impact is not very pronounced, while *cob*, *cox1*, and *nad5* contribute the most to this topology (see coef column: large absolute value means large effect, while negative coefficients tend to break the topology). All these findings are coherent with the frequency of occurrences of each gene in the choice of Topology 1: *nad5*, *cox1*, and *nad5* were present in 12,642 computations leading to this topology (77.07%), while only 8,685 computations with *rrnS*, *nad3*, and *nad1* have led to this topology (53%), as described in Table V.

Further investigations of the role of each sequence and their effects on each topologies are provided in Tables VII, VIII, and IX of supplementary data, which contain the results of the dummy logit regression test for Topologies 0, 2, and 3 respectively.

## IV. CONCLUSION

Deep investigation of the molecular phylogeny of the *Taenia* genus has been performed in this paper. 14 coding sequences, taken from mitochondrial genomes, have been considered for maximum likelihood phylogenetic reconstruction. As the obtained tree was not satisfactorily supported, each combination from 1 to 14 genes has been further investigated, leading to 16,384 trees representing 131 topologies. Four close topologies were then isolated whose differences are located in the position of *T. hydatigena* and the sister relationship between *T. laticollis* and *T. pisiformis*. Using the logit model we have finally proven that Topology 1 was the most probable one and have emphasized the negative role of some genes for that phylogeny.

In future work, the authors intend to use LASSO test for regressing the bootstrap on the genes. Furthermore, we will

investigate the phylogeny of *Echinococcus* using a similar approach. Indeed, there is no general agreement regarding the phylogeny of this genus. In particular, some species were discovered to have contradictory positions in the available literature. All the possible combinations of the 12 mitochondrial genes, plus *rrnL* and *rrnS* and also 5 nuclear genes, will be considered, leading to the production of 43,796 phylogenetic trees. Their topologies will be compared, and the influence of each gene on these topologies will be rigorously measured, in order to determine the most probable phylogenetic tree of this species. Finally, the phylogeny of the class *Eucestoda* will be investigated using a similar approach.

*All computations have been performed using the Mésocentre de Calcul de Franche-Comté facilities.*

## REFERENCES

- [1] Bassam AlKindy, Huda Al-Nayyef, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques Bahi. Improved core genes prediction for constructing well-supported phylogenetic trees in large sets of plant species. In *Bioinformatics and Biomedical Engineering*, volume 9043, pages 379 – 390, Granada, Spain, apr 2015. Springer.
- [2] Bassam AlKindy, Bashar Al-Nuaimi, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Reem Alsraraj, and Laurent Philippe. Binary particle swarm optimization versus hybrid genetic algorithm for inferring well supported phylogenetic trees. *Computational Intelligence Methods for Bioinformatics and Biostatistics*, 9874:165–179, 2016. Revised and extended journal version of the CIBB2015 conference.
- [3] Bassam AlKindy, Jean-François Couchot, Christophe Guyeux, Arnaud Mouly, Michel Salomon, and Jacques M. Bahi. Finding the core-genes of chloroplasts. *Journal of Bioscience, Biochemistry, and Bioinformatics*, 4(5):357–364, 2014. Journal version of ICBBS14 conference.
- [4] J. William O. Ballard and Michael C. Whitlock. The incomplete natural history of mitochondria. *Molecular Ecology*, 13(4):729–744, 2004.
- [5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Res.*, 37(Database issue):26–31, Jan 2009.
- [6] R. C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1), August 2004.
- [7] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791, 1985.
- [8] Nicolas Fiorini, Vincent Lefort, François Chevenet, Vincent Berry, and Anne-Muriel A Chifolleau. CompPhy: a web-based collaborative platform for comparing phylogenies. *BMC evolutionary biology*, 14(1):253, 2014.
- [9] S. Guindon, JF Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phylml 3.0. *Systematic Biology*, 59(3):307–321, 2010.
- [10] Michael Hardman and Lotta M. Hardman. Comparison of the phylogenetic performance of neodermatan mitochondrial protein-coding genes. *Zoologica Scripta*, 35(6):655–665, 2006.
- [11] E.P. Hoberg. Phylogeny of taenia: Species definitions and origins of human parasites. *Parasitol Int*, 55 Suppl, 2006.
- [12] E.P. Hoberg, A. Jones, R.L. Rausch, K.S. Eom, and S.L. Gardner. A phylogenetic hypothesis for species of the genus taenia (eucestoda : Taeniidae). *J Parasitol*, 86(1):89–98, 2000.
- [13] Eric P. Hoberg, Nancy L. Alkire, Alan D. Queiroz, and Arlene Jones. Out of Africa : origins of the Taenia tapeworms in humans. (September 2000), 2001.
- [14] Jenny Knapp, Minoru Nakao, Tetsuya Yanagida, Munehiro Okamoto, Urmaz Saarma, Antti Lavikainen, and Akira Ito. Phylogenetic relationships within echinococcus and taenia tapeworms (cestoda: Taeniidae): An inference from nuclear protein-coding genes. *Mol Phylogenet Evol*, 2011.
- [15] Nicolas Lartillot, Thomas Lepage, and Samuel Blanquart. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288, September 2009.
- [16] Antti Lavikainen, Voitto Haukialmi, Markus J. Lehtinen, Sauli Laaksonen, Sauli Holmström, Marja Isomursu, Antti Oksanen, and Seppo Meri. Mitochondrial {DNA} data reveal cryptic species within taenia krabbei. *Parasitology International*, 59(2):290 – 293, 2010.
- [17] Minoru Nakao, Antti Lavikainen, Takashi Iwaki, Voitto Haukialmi, Sergey Konyaev, Yuzaburo Oku, Munehiro Okamoto, and Akira Ito. Molecular phylogeny of the genus taenia (cestoda: Taeniidae): proposals for the resurrection of hydatigera lamarck, 1816 and the creation of a new genus versteria. *Int J Parasitol*, 2013.
- [18] Minoru Nakao, Tetsuya Yanagida, Munehiro Okamoto, Jenny Knapp, Agathe Nkouawa, Yasuhito Sako, and Akira Ito. State-of-the-art echinococcus and taenia: Phylogenetic taxonomy of human-pathogenic tapeworms and its application to molecular diagnosis. *Infection, Genetics and Evolution*, 10(4):444 – 452, 2010.
- [19] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, September 2000.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [21] Vincent Ranwez, Alexis Criscuolo, and Emmanuel J.P. Douzery. Supertriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics*, 26(12):i115–i123, 2010.
- [22] Larry S Roberts and John Janovy. *Gerald D. Schmidt & Larry S. Roberts' Foundations of Parasitology*. McGraw-Hill Higher Education Boston, 2005.
- [23] Ros Stamatakis and Thomas Ludwig. Raxml-omp: An efficient program for phylogenetic inference on smps. In *In Proc. of PaCT05*, pages 288–302, 2005.
- [24] Stacia K. Wyman, Robert K. Jansen, and Jeffrey L. Boore. Automatic annotation of organellar genomes with dogma. *Bioinformatics*, 20(17):3252–3255, 2004.

## V. APPENDICES

TABLE VII: Dummy logit regression results for Topology 0

	coef	std err	z	P >  z	[95.0% Conf. Int.]
<i>atp6</i>	-1.0959	0.069	-15.901	0.000	[-1.231, -0.961]
<i>cob</i>	-1.7306	0.073	-23.593	0.000	[-1.874, -1.587]
<i>cox1</i>	0.233	0.066	3.535	0.000	[0.104, 0.362]
<i>cox2</i>	-0.033	0.066	-0.501	0.616	[-0.162, 0.096]
<i>cox3</i>	1.4431	0.071	20.327	0.000	[1.304, 1.582]
<i>nad1</i>	-2.6491	0.082	-32.159	0.000	[-2.811, -2.488]
<i>nad2</i>	-5.767	0.151	-38.171	0.000	[-6.063, -5.471]
<i>nad3</i>	-0.0797	0.066	-1.211	0.226	[-0.209, 0.049]
<i>nad4</i>	2.4925	0.08	31.017	0.000	[2.335, 2.650]
<i>nad4l</i>	-0.0296	0.066	-0.449	0.653	[-0.159, 0.099]
<i>nad5</i>	-2.5196	0.081	-31.123	0.000	[-2.678, -2.361]
<i>nad6</i>	-1.0355	0.069	-15.097	0.000	[-1.170, -0.901]
<i>rrnL</i>	-1.403	0.071	-19.803	0.000	[-1.542, -1.264]
<i>rrnS</i>	2.2175	0.078	28.594	0.000	[2.066, 2.370]

TABLE VIII: Dummy logit regression results for Topology 2

	coef	std err	z	P >  z	[95.0% Conf. Int.]
<i>atp6</i>	0.3534	0.055	6.479	0.000	[0.247, 0.460]
<i>cob</i>	-0.3845	0.055	-7.042	0.000	[-0.492, -0.277]
<i>cox1</i>	-1.5321	0.059	-25.903	0.000	[-1.648, -1.416]
<i>cox2</i>	-1.0754	0.057	-18.956	0.000	[-1.187, -0.964]
<i>cox3</i>	-1.7371	0.06	-28.729	0.000	[-1.856, -1.619]
<i>nad1</i>	2.305	0.065	35.601	0.000	[2.178, 2.432]
<i>nad2</i>	3.6525	0.078	46.902	0.000	[3.500, 3.805]
<i>nad3</i>	0.8116	0.056	14.577	0.000	[0.702, 0.921]
<i>nad4</i>	-1.4916	0.059	-25.323	0.000	[-1.607, -1.376]
<i>nad4l</i>	-1.641	0.06	-27.427	0.000	[-1.758, -1.524]
<i>nad5</i>	-3.4317	0.075	-45.481	0.000	[-3.580, -3.284]
<i>nad6</i>	0.2363	0.054	4.343	0.000	[0.130, 0.343]
<i>rrnL</i>	-0.7259	0.055	-13.1	0.000	[-0.834, -0.617]
<i>rrnS</i>	-1.3261	0.058	-22.878	0.000	[-1.440, -1.213]

TABLE IX: Dummy logit regression results for Topology 3

	<b>coef</b>	<b>std err</b>	<b>z</b>	$P >  z $	<b>[95.0% Conf. Int.]</b>
<i>atp6</i>	-0.9133	0.081	-11.331	0.000	[-1.071, -0.755]
<i>cob</i>	-1.3941	0.085	-16.49	0.000	[-1.560, -1.228]
<i>cox1</i>	-0.1349	0.078	-1.739	0.082	[-0.287, 0.017]
<i>cox2</i>	0.8051	0.08	10.12	0.000	[0.649, 0.961]
<i>cox3</i>	-0.7384	0.08	-9.278	0.000	[-0.894, -0.582]
<i>nad1</i>	-2.0258	0.092	-22.074	0.000	[-2.206, -1.846]
<i>nad2</i>	-0.2326	0.078	-2.992	0.003	[-0.385, -0.080]
<i>nad3</i>	-0.7489	0.08	-9.406	0.000	[-0.905, -0.593]
<i>nad4</i>	-3.5617	0.131	-27.208	0.000	[-3.818, -3.305]
<i>nad4l</i>	0.4031	0.078	5.168	0.000	[0.250, 0.556]
<i>nad5</i>	2.1308	0.092	23.219	0.000	[1.951, 2.311]
<i>nad6</i>	-1.1314	0.082	-13.766	0.000	[-1.292, -0.970]
<i>rrnL</i>	-3.8926	0.146	-26.68	0.000	[-4.179, -3.607]
<i>rrnS</i>	0.2623	0.078	3.376	0.001	[0.110, 0.415]